ED 154 022 TM 007 035

AUTHOR Webster, Raymond E.; Bates, Herman M., III
TITLE Federal Legislation Defining Learning Disabilities and Biased IQ Scores.
SPONS AGENCY Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C.
PUB DATE Aug 77
CONTRACT 443-CH-60166
GRANT G007605223
NOTE 13p.; Paper presented at the Annual Meeting of the American Psychological Association (85th, San Francisco, California, August 26-30, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Age Differences; Educational Legislation; Elementary Secondary Education; Federal Legislation; *Intelligence Quotient; *Intelligence Tests; *Learning Disabilities; Predictive Validity; Student Placement; Test Bias; *Test Reliability; *True Scores
IDENTIFIERS *Test Retest Reliability

ABSTRACT

Public Law 94-142 mandates the identification and placement of learning disabled children based primarily on the measurement of intelligence. It is, therefore, the responsibility of educational psychologists to use standardized intelligence tests appropriately to accurately and objectively assess a childs' intellectual potential and ability. Three general assumptions underlying the measurement of intelligence are discussed: (1) that intelligence is both measurable and quantifiable; (2) that it is distributed according to a normal curve; and (3) that intelligence remains constant over time. The constancy of intelligence level and the stability of its measurement (test reliability) are important issues. Test reliability may be established according to internal consistency; correlation with equivalent or parallel tests; or, most importantly, consistency of measurement over time. Data on the stability of certain intelligence tests over time are presented. These limited data suggest that the scores of young children are less stable than of older children and adults; that stability decreases as the length of the test-retest interval increases; and that children with various disabilities exhibit more test-retest variation. A formula is presented for estimating an unbiased true score. More research is needed investigating fair methods of placing learning disabled children. (GDC)

FEDERAL LEGISLATION DEFINING LEARNING DISABILITIES AND BIASED IQ SCORES*

A paper selected for presentation at the American Psychological Association Annual Convention, Division #16 "Poster Session", San Francisco, August, 1977.

Raymond E. Webster

Herman M. Bates, III

Department of Educational Psychology
University of Connecticut
Box U-7
Storrs, Connecticut 06268

*Preparation of this paper was aided in part by a grant from the Bureau of Education for the Handicapped, USOE, "Learning Disabilities in Mathematics: A Curriculum Design for Upper Grades", G007605223, #443CH60166, under the direction of J. F. Cawley, Project Director, University of Connecticut, Storrs, Connecticut 06268

FEDERAL LEGISLATION DEFINING LEARNING
DISABILITIES AND BIASED IQ SCORES

In recent years the placement of children solely on the basis of an IQ score, or a battery of tests purportedly measuring IQ, is a measurement procedure that has been seriously questioned. Placements based upon this prediction model are generally administratively beneficial. However, incorrect decisions can often be damaging to the individual child involved.

Section 5 of the Proposed Rules on Specific Learning Disabilities issued by the Office of Education of the Department of Health, Education and Welfare on November 29, 1976 (referred to as Public Law 94-142) coerces the prediction and placement of children primarily on the basis of measurement of intelligence.

As such, it is the responsibility of school and educational psycholgists to use standardized intelligence tests in the most appropriate manner to accurately and objectively assess a child's intellectual potential and ability.

There are a number of theoretical and general assumptions underlying the measurement of intelligence. Perhaps the most obvious is that the global entity intelligence is both measurable and quantifiable. Aside from the problems caused by the lack of a common operational definition for intelligence, this assumption tends to ignore individual fluctuations in capacity to deal with day to day situations, individual mood swings, transient physical disabilities, and most importantly, socio-economic and cultural background differences.

A second assumption, and one that can neither be proven nor disproven, is that intelligence within the general population is normally distributed according to the Gaussian Curve. Dingman and Tarjan (1960) have collected data that seriously questions this assumption. They calculated the expected number of people at various low IQ levels according to the normal curve. They then compared these estimates to the population of the United States, 210 million. Large discrepancies were found between predicted estimates from the Gaussian Curve and actual numbers. For example, in the 0 to 20 IQ range the predicted number of cases is 57. Dingman and Tarjan found the actual prevalence of cases in this IQ range to be 104,935.

Another extremely important assumption, and one that underlies both the measurement of intelligence and the prediction-placement model, is that intelligence remains constant over time. Certainly, this topic has been a recurring issue in recent years. Schaie (1974) has argued that the results from research

> "combining the rigor of the scientific method with an address to problems that may indeed be of social consequence" now indicate that "a presumed decline in adult intelligence is at best a methodological artifact and at worst a popular misunderstanding of the relation between individual development and sociocultural change." (p. 802)

Baltes and Schaie (1974) and Schaie, Labouvie-Vief and Barrett (1973) suggest that the decrement hypothesis is based on poor psychometric or biomedical models. This viewpoint has been hotly contested by Horn and Donaldson (1976, 1977). Although the crux of this argument has focused on the decline of intelligence related to old age, the constancy position of intelligence may be criticized because it ignores individual variability related to the physical and psychological development of especially the late-developing child. Binet himself (quoted by Skeels and Dye, 1939) commented that:

> Some recent philosophers appear to have given their moral support to the deplorable verdict that the intelligence of an individual is a fixed quantity, a quantity which cannot be augmented. We must protest and act against this brutal pessimism. We shall endeavour to show that it has no foundation whatsoever ....A child's mind is like a field for which an expert farmer has advised a change in the method of cultivating, with the result that in place of desert land we now have a harvest. It is in this particular sense, the only one that is significant, that we say that the

> intelligence of children may be increased. One increases that which constitutes the intelligence of a school child; namely the capacity to learn, to improve with instruction.

Despite these arguments, child development textbooks continue to implicitly promulgate the idea that intelligence is a temporally stabilized phenomenon. Clarke and Clarke (1953) examined a paper by Nemzek (1933) and another by Thorndike (1940). Together the two papers reviewed a total of 359 studies of intelligence measurement carried out before 1940. Clarke and Clarke (1953) concluded that (1) the predictive value of the IQ as measured by test-retest correlations decreases as the interval between testings increases; (2) although the average IQ of the population may not significantly change, some individuals exhibit significant variability in IQ measurement; (3) intelligence tests given to children before entering school have little value in predicting later achievement; and (4) mental assessments for infants have no predictive relevance in later years. Again, the constancy of IQ measurement is questioned.

A complementary issue to the question of IQ constancy is the question of measurement constancy. This phenomenon is referred to as test-retest reliability, or stability of measurement over time.

The basic theoretical issue in the stability of intelligence assessment is the notion of the existence and measurability of a "true" score. This "true" score is associated with an individual's obtained score on an IQ test. The classical theory of reliability offers three alternative ways of defining a "true" score (Lord & Novick, 1968, pp. 28-29). The first notion, referred to by Sutcliffe (1965) as the Platonic "true" score, suggests that a "true" score exists for each observation. This score is not observable because it is obscured by measurement errors. A major objection to this position is that the "true" score can never be measured or quantified (Thorndike, 1964). A more widely accepted position is that the "true" score is a probabilistic entity. In principle, if one were to make an infinite number of observations of some attribute, the mean number of these observations would converge toward a constant or "true" score. The final definition, and perhaps the most mathematically rigorous, suggests that "....corresponding true and error scores are uncorrelated and that error scores on different measures are also uncorrelated...." (Lord and Novick, 1968, p. 29). The binding element of the three definitions is the existence of an exact, single "true" score. Accurate measurement of this score represents perfect reliability. As the discrepancy between the "true" score and an obtained score increases the instrument's reliability decreases.

There are three ways in which to measure or establish reliability. The most common method refers to the internal consistency of an instrument. This may be computed by using the procedures outlined by Kuder and Richardson (1937). Reliability can also be established through the process of equivalence, where the parallel form of a test is administered at the same time or a short time after, and then correlated with the instrument in question. The final form of reliability, and perhaps the most important when dealing with intelligence measurement, refers to the consistency or stability of measurement over time. In this procedure the same form of an instrument is administered two (or more) times to the same sample population separated by varying intervals of time. In the case of intelligence tests, this period of time should extend over a minimum of months, and ideally, years. This is especially important if one subscribes to the IQ constancy paradigm, a position implicitly accepted in most public schools, mental health clinics and hospitals.

The initial reports of reliability for the WISC (Wechsler, 1949) were established through a split-half reliability procedure, a measure of internal consistency and not of stability of measurement over time. More recently for the WISC-R (Wechsler, 1974) stability coefficients for periods of three to five weeks are reported. However, the instrument's stability over longer periods of time has not been systematically investigated. The 1960 revision of the Stanford-Binet reports no test-retest reliability coefficients. There remains a serious lack of research examining the stability of all general intelligence measures over extended intervals of time (three months and longer), with various diagnostic groups of children. Table 1 contains the results of 12 studies examining the stability of the WPPSI, WISC, WAIS, and the Stanford-Binet forms L and M found through a review of the research literature. No studies examining the stability of WISC-R scores were found.

TABLE 1 ABOUT HERE

Inspection of the table shows that there has been a paucity of research investigating the stability of intelligence tests over time. Thus, it is difficult to determine general trends and patterns throughout the research. However, it may be concluded that (1) the scores of young children are less stable than those of older children and adults; (2) as the length of the test-retest interval increases, the stability coefficient decreases; and (3) children who exhibit different types of disabilities show more test-retest variation. Eysenck (1953) has shown that the test-retest correlation coefficient for large groups decreases steadily at the rate of about 0.04 a year. If one assumes that the average test-retest reliability coefficient after immediate retesting is around .90 (Thorndike, 1933, 1940) then after seven years a test-retest coefficient of around .62 is expected. Subsequently, only 38.44% of the total variation is attributable specifically to the intelligence measure. This suggests that if intelligence scores are to be used for prediction and placement then children with specific disabilities should be frequently re-evaluated. Frequent re-assessment is even more important for the younger disabled child because IQ estimates may be confounded by both developmental factors and the nature and extent of the disability. There is a definite need for more systematic research on the stability of IQ scores as a function of chronological age and of emotional and learning disabilities.

As the test-retest reliability of an instrument decreases, the estimated true score shows a greater tendency to regress toward the mean. The most serious implications of this regression are for extreme scores or those that lie farthest from the mean.

The obtained score on an intelligence test with less than perfect reliability represents an outwardly biased estimate of the child's "true" score and ability. Thus, for low IQ scores, the estimated "true" scores tend to be higher, and for high IQ scores they tend to be lower.

An unbiased estimate or "true" score for a child can be computed by using the following formula suggested by Nunnally (1967, pp. 220-221):

$t' = r_{x1x2}X$ where,

t' = the estimated unbiased deviation score;

r_{x1x2} = the reliability of the test used; and

X = the deviation score (the obtained score minus the mean)

The "true" score or best unbiased estimate of IQ = mean of the test instrument (usually 100) + t'.

This formula yields the best unbiased estimate about which confidence intervals should be established. This unbiased estimate is also the fairest number (to the child) to be used in the optional "severe discrepancy" formula or as the basis in determining if a child meets the 50% discrepancy between ability and achievement.

The following example shows how the above formula might be used. Coleman (1963) found that the test-retest reliability coefficient for the Full Scale IQ score of the WISC for learning disabled children of age 7.5 years is .77, with a standard error of measurement of 8.61. A child with an obtained IQ score of 80 would have a deviation score (X) of -20, and an estimated unbiased "true" IQ score of 84.6, or rounded to 85. The confidence interval for the 95% level of confidence is 85$\pm$ 16.88. The asymmetrical nature of the confidence interval around the observed IQ score of 80 serves as a reminder of the biased nature of the obtained score as a function of the error intrinsic to the instrument.

Concluding Remarks

There is a great need for continued systematic research investigating the stability of intelligence test measurements for various diagnostic groups. The procedures used to place children into special classes, programs or institutional settings should be modified until empirical research confirms and/or defines the parameters of predictive efficacy of these tests. An unbiased estimate of an IQ score should probably be computed and used in these modified placement procedures in an effort to be fair to each child.

To clarify the definition of learning disability, the federal government has perpetuated a measurement dilemma. The optional formula to predict whether a child is severely discrepant between ability and achievement demands the determination of the child's "true" IQ. This situation becomes even more complicated when one considers that at least two measures of intelligence must be used to determine the IQ score.

It is proposed that the best unbiased estimation be computed and used as a fairer estimate of a child's true intellectual ability and potential.

Table 1

Stability Coefficients for Widely Used Intelligence Measures

Instrument	Sample Characteristics	Test-Retest Interval	Stability r	Authors
WISC	n = 39, IQ range 40-79 Institutionalized MR Age Range 11-0 : 14-11	3 - 4 mos.	Verbal = .92 Perfor. = .89 FSIQ = .95	Throne, Schulman & Kasper (1962)
WISC	n = 26, anti-social and habit disordered Age: 7-7:15-1, $\bar{X}$=12-1	6 mos.	Verbal = .81 Perfor. = .73 FSIQ = .80	Turner, Mathews & Rachman (1967)
WISC	mentally retarded		Verbal = .48 Perfor. = .78 FSIQ = .88	Friedman (1970)
WISC	n = 24 males; learning disabled $\bar{X}$ IQ - 102		Verbal = .62 Perfor. = .81 FSIQ = .77	Coleman (1963)
WISC	n = 21: 11 males and 10 females; emotionally disturbed children	2 to 15 mos. $\bar{X}$ = 7.8 mos.	Verbal = .819 Perfor. = .508 FSIQ = .834	Tigay & Kempler (1971)
WISC Stanford-Binet	IQ = 55 to 75 children had no sensory or behavioral problems	3 yrs.	Verbal = .70 Perfor. = .73 FSIQ = .76 SBIQ = .79	Walker & Gross (1970)
WISC Stanford-Binet	n = 60; "normal" children; tested in 5th and 9th grades	4 yrs	Verbal = .77 Perfor. = .74 FSIQ = .77 SBIQ = .78	Gehman & Matyas (1956)
Stanford-Binet	n = 182 mentally retarded students IQ = 42 to 89 age range: 6 to 15 yrs.		SBIQ = .93	Collman & Newlyn (1958)
Stanford-Binet	"normal" subjects aged 2 yrs. to 18 yrs.	3 yrs	age range r 2-5: .32 $\pm$ .06 3-6: .57 $\pm$ .05 4-7: .59 $\pm$ .04 5-8: .70 7-10: .78 9-12/13: .85 14/15-18: .79	Honzik, MacFarlane & Allen (1948)
Stanford-Binet	n = 111 "normals" first tested as children	10 yrs. 15 yrs. 25 yrs.	SBIQ 1931 to 41 .65 1941 to 56 .85 1931 to 56 .59	Bradway & Thompson (1962)
WISC WAIS	WISC: n = 46 mentally retarded children; WAIS: n = 130 mentally retarded adults	WISC $\bar{X}$ = 33 mos. WAIS $\bar{X}$ = 29.5 mos.	WISC Verbal = .70 Perfor. = .72 FSIQ = .81 WAIS Verbal = .87 Perfor. = .92 FSIQ = .88	Rosen, Stallings, Floor & Nowakiwska (1968)
WPPSI	n = 50 "normal" 5 yr. olds	3 mos.	FSIQ = .92	Oldridge & Allison (1968)

References

Baltes, P. and Schaie, K. The myth of the twilight years. Psychology Today, 1974 (March), 35-38; 4(

Bradway, K. P. and Thompson, C. W. Intelligence at adulthood: a twenty-five year followup. Journal of Educational Psychology, 1962, 53, 1-14.

Clarke, A. D. B. and Clarke, A. M. How constant is the IQ? Lancet, 1953, ii, 877, 880.

Coleman, J. C. Stability of intelligence test scores in learning disorders. Journal of Clinical Psychology, 1963, 19, 295-298.

Dingman, H. and Tarjan, G. Mental retardation and the social distribution curve. American Journal of Mental Deficiency, 1960, 64, 991-994.

Eysenck, H. J. Uses and Abuses of Psychology. Harmondsworth: Penguin Books, 1953.

Friedman, R. The reliability of the Wechsler Intelligence Scale for Children in a group of mentally retarded children. Journal of Clinical Psychology, 1970, 26, 181-182.

Gehman, I. H., and Matyas, R. P. Stability of the WISC and Binet tests. Journal of Consulting Psychology, 1956, 20, 150-152.

Honzik, M. P., MacFarlane, J. W., and Allen L. The stability of mental test performance between two and eighteen years. Journal of Experimental Education, 1948, 17, 309-324.

Horn, J. and Donaldson, G. On the myth of intellectual decline in adulthood. American Psychologist, 1976, 31, 701-719.

Horn, J. and Donaldson, G. Faith is not enough: a response to the Baltes-Schaie claim that intelligence does not wane. American Psychologist, 1977, 32(5) 369-373.

Kuder, G. F. and Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160.

Lord, F. and Novick, M. Statistical Series of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968.

Nemzeck, C. L. The constancy of IQ. Psychological Bulletin, 1933, 30, 143-168.

Nunnally, J. C. Psychometric Theory. New York: McGraw-Hill Book Company, 1967.

Oldridge, O. A. and Allison, E. E. Review of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI). Journal of Educational Measurement, 1968, 5, 347-348.

Rosen, M., Stallings, L., Floor, L., and Nowakiwska, M. Reliability and stability of Wechsler IQ scores for institutionalized mental subnormals. American Journal of Mental Deficiency, 1968, 7[illegible], 218-225.

Schaie, K. W. Translations in gerontology - from lab to life: intellectual function. American Psychologist, 1974, 29, 802-807.

Schaie, K. and Labouvie-Vief, G. and Barrett, T. Selective attention effects in a fourteen-year study of adult intelligence. Journal of Gerontology, 1973, 28, 328-334.

Skeels, H. M. and Dye, H. B. A study of the effects of differential stimulation on mentally retarded children. American Association of Mental Deficiency, 1939, 44, 114-1[illegible].

Thorndike, R. L. The effect of the interval between test and retest on the constancy of I.Q. Journal of Educational Psychology, 1933, 24, 543-549.

Thorndike, R. L. "Constancy" of IQ. Psychological Bulletin, 1940, 37, 167-186.

Throne, F. M., Kaspar, J. C., and Schulman, J. L. Reliability and stability of the Wechsler Intelligence Scale for Children for a group of mentally retarded boys. American Journal of Mental Deficiency, 1962, 67, 455-457.

Tigay, B., and Kempler, H. L. Stability of WISC scores of children hospitalized for emotional disturbance. Perceptual & Motor Skills, 1971, 32, 487-490.

Turner, R. K., Mathews, A., and Rachman, S. The stability of the WISC in a psychiatric group. British Journal of Educational Psychology, 1967, 37, 194-[illegible].

Walker, [illegible], and Gross, F. L. IQ stability among educable mentally retarded children. Training School Bulletin, 1970, 66, 181-187.

Wechsler, D. Manual for the Wechsler Intelligence Scale for Children. New York: Psychological Corporation, 1949.

Wechsler, D. Manual for the Wechsler Intelligence Scale for Children - Revised. New York: Psychological Corporation, 1974.